Curators' Evaluation of WAS Release 4

October 24, 2007

Prepared by:

Kathleen R. Murray
krmurray@unt.edu

University of North Texas

Contents

1 Introduction 2

2 Sites 3

3 Capture Results 5

4 Collections 8

5 Overall Reactions 10

1 Introduction

The Web Archiving Service, WAS Release 4, was available to the project's curators from September 18, 2007 – October 12, 2007. During this time project curators trialed the service and, subsequently, 17 curators submitted a web-based survey evaluation of their experience. The survey consisted of 25 questions, divided into four sections: Sites, Capture Results, Collections, and Overall Release.

Overall, curators were very satisfied with this WAS release, as one curator commented: "I can finally envision how I would use it [the WAS] to build collections." Another curator wrote: "I'm very optimistic I will be able to make this work for capturing and organizing my collections."

As with WAS release 2/3, opinions about the benefit and usefulness of new and enhanced features in this release were not unanimous. However, several curators welcomed the improvements in overall workflow, available help, and the speed of captures. The new features curators mentioned as some of the best in this release were the bookmarklet, which allowed curators to define sites as they browsed the Web, the RSS feed, which provided the capture completion status of a curator's most recent captures, and the ability to compare two different captures of the same site, which one curator thought might be "a backbone feature for WAS."

This release appeared to effectively resolve three problems in WAS release 2/3: (1) estimating max time capture settings for sites, (2) unexpected search results and lack of understanding about what content was being searched, and (3) discovery and display of discrete files of publications or documents, especially PDF files. A fourth problem curators reported in WAS release 2/3 was partially addressed; while no curators reported receiving either an error message or a "Not in Archive" message for files listed in the captured files list, one curator reported they continued to be unable to display all files from captured sites. A fifth problem, display of the "home page" for seed URLs, remains a problem with 71% of curators (n=12) reporting they had difficulty displaying the content of seed URLs.

The major areas that were either confusing or problematic in this release were:

1. Receiving an error message when attempting to search collections (Message: 'There was an error getting the index for searching or listing files. This index may be very large and is taking a long time with its initial loading.')
2. Display of the "home page" for seed URLs
3. Installation of the bookmarklet and RSS feed at campuses whose IT policies prohibit end users from installing programs/tools such as these

One curator reported the following two problems. These are included here because of their seriousness and are described in the referenced sections of the report.

1. Inconsistency between reports of site captures and captured content (Section 3.1.)
2. Links from captured files leading to live web sites outside of the archive with no alert (Section 3.2.)

The remainder of this report summarizes the feedback received from the curators and provides details to illustrate the areas that were either confusing or problematic to them. Their recommendations as well some considerations for future development are also included.

2 Sites

2.1 Descriptive Metadata

Three metadata elements were available for describing each site curators defined: creator, subjects, and coverage (geographic). As shown in Table 1 the majority of curators anticipated that all three would be very useful for describing the sites in their collections. Only two curators thought creator would not be useful in describing their sites and one curator did not anticipate that geographic coverage would be useful in describing their sites.

	N	Very Useful		Somewhat Useful		Not Useful		Not Sure	
Metadata Element		**#**	**%**	**#**	**%**	**#**	**%**	**#**	**%**
Creator	17	11	65%	3	18%	2	12%	1	6%
Subjects	17	9	53%	7	41%	0	0%	1	6%
Coverage (geographic)	17	12	71%	3	18%	1	6%	1	6%

Table 1. Usefulness of Metadata Elements

Comments from Curators:

- I didn't really know what to put as subjects--will be useful in the future but I didn't know how to be consistent in language use.
- I did not add any subjects at this time because we are not at the stage where we have a metadata scheme that would be used in a real application, however, it would be useful.
- I found myself only adding basic metadata, if any, due to lack of time. However, the flexibility to go back and edit site records will enable me to add more at a later time.

2.2 Bookmarklet

A bookmarklet that allowed curators to define sites as they browsed the Web was available for installation. Most of the curators (11 of 17) had no difficulty installing the bookmarklet. Five curators did not attempt to install it and one curator did have some difficulty. For some curators, their institution's IT policies prohibit them from installing anything, including applications and tools, on their computers.

Comments from Curators:

- Installing the bookmarklet will be difficult/impossible at work for those of us whose computers are so tied down that we can't install anything w/o systems help.
- Our IT environment does not allow users to install things on their computers.

Of the 12 curators who installed the bookmarklet, eight (67%) indicated that it would definitely be useful to them for adding sites they want to capture and four (33%) indicated it would be somewhat useful to them.

Comments from Curators:

- Really like the bookmarklet
- The bookmarklet is great - I absolutely love it!

2.3 RSS Feed

Curators had the option to install an RSS feed that provided the capture completion status of their most recent captures without curators needing to be logged into the Web Archiving Service. Seven curators installed the RSS feed with no difficulty. Only one curator who attempted to install it was unable to do so. Of the seven curators who did install the feed, five rated it as definitely of benefit to them or as somewhat of benefit to them. Two curators did not think it was of benefit to them (Table 2).

N=17	**Benefit Rating**	**#**	**%**
	Definitely	3	18%
	Somewhat	2	12%
	Not Really	2	12%
	Not Tried	10	59%

Table 2. Benefit of RSS Feed

Over half of the curators (10 of 17) did not use this feature. As with installation of the bookmarklet, some institutions' IT policies prohibit users from installing anything on their computers. For one curator the timing of the WAS release was not optimal and the curator did not have enough time to trial either the RSS feed or the bookmarklet.

Comments from Curators:

- I would have liked to have tried the RSS feed, but it would have required a call to the [IT] help desk to have them install.

2.4 Recommendations and Considerations

Recommendations from Curators:

- Adding the name of the collection[s] to which each site belongs to the list of sites would be useful
- The Bookmarklet would be more helpful if it remembered the curator's name and password
- Add a notes field where curators could record any kind of information about the capture, other than 'descriptive' kinds of things. For example - before capturing sites - I check for robots, forms that have to be filled in to access info, use of flash, etc.; or I might choose to capture just a directory or host - and I have a specific reason for doing so and I'd like to record this information so I can keep track of it. Later on, I might have other kinds of notes when building collections that I'd like to record as well.
- It would be nice to have a way to get around robot exclusions.

Considerations:

- The project curators are responsible for collections largely in the areas of government and political information, for which geographic coverage is of importance. A wider diversity of collection focus among future curators might yield more variance in the usefulness of this metadata element as well as suggest other elements of importance.
- Consider adding a note in the documentation for the bookmarklet and the RSS feed advising curators that they may need to contact their IT organization for installation of these tools.

3 Capture Results

3.1 Overview Report

Most curators indicated the Overview Report was helpful in evaluating their completed captures, with 56% (*n*=9) rating it very helpful and 25% (*n*=4) rating it somewhat helpful. However, unlike the previous WAS release in which all curators (*N*=17) rated the report as either very helpful or somewhat helpful, two curators did not find the report helpful in evaluating their completed captures using WAS release 4 (Table 3).

N=16		#	%
	Very Helpful	9	56%
	Somewhat Helpful	4	25%
	Not Helpful	2	13%
	Not Used	1	6%

Table 3. Helpfulness of Overview Report

One curator reported that the Overview Report for only one of their four sites accurately reflected what was captured. This curator discovered inconsistencies between the Report data and the captured content. These inconsistencies are listed below.

Problems Reported by Curator:

- Two sites, each captured twice and containing no robots.txt files:
  - Seemed to go through the full crawl process and generated reports for mimetypes and hosts. Yet none of the actual files were captured -- there was no captured content.
  - The capture results also listed capture times which seemed random. They weren't even just wrong; they were hours apart from each other even though the crawls were done minutes apart.
- A third site . . . listed no robots.txt file in the Overview reports, but did list such a file in the file list.

3.2 Content Display

Overall, curators were either very satisfied (*n*=8; 47%) or somewhat satisfied (*n*=6; 35%) with the display of captured content. Three curators were not satisfied with the display of content. One curator reported that hyperlinks within captured content led to live Web sites. Another reported that some files could not be displayed at all.

Figure 1 compares curators' satisfaction in regard to the display of captured content between WAS Release 2/3 and Release 4. Curators' overall satisfaction increased with Release 4.

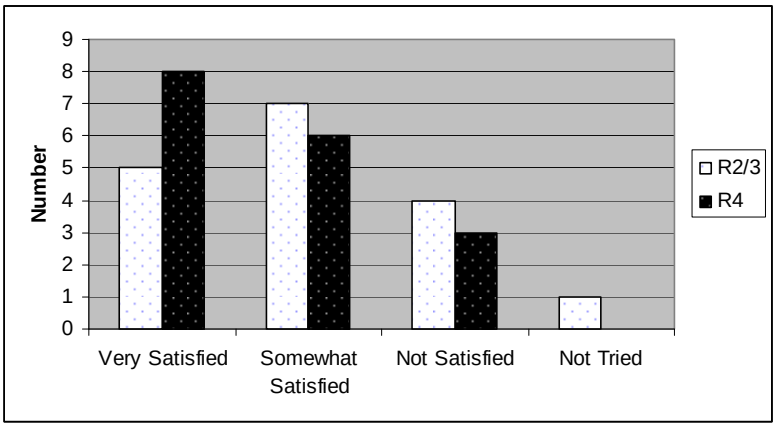

Figure 1. Comparison of Overall Satisfaction with Displayed Content between WAS Release 2/3 and Release 4

Most (*n*=12; 71%) of the curators had difficulty displaying the content of seed URLs. One curator commented that it was simply not evident how to do this.

As shown in Table 4, the majority (*n*=10; 59%) of curators easily displayed content from different captures of the same site or file. However, three curators had difficulty doing this.

N=17		#	%
	Yes	10	59%
	No	3	18%
	Not Tried	4	24%

Table 4. Multiple Site Capture Display

Ten curators (59%) could anticipate that the ability to add comments from the detailed record display for files would be either very useful (*n*=6) or somewhat useful (*n*=4) for the files in their collections. Seven curators (41%) were not sure if this feature would be useful for their collections.

Comments from Curators:

- It seemed hard to find the descriptive metadata. (Note: Perhaps the curator meant the detailed record display?)
- The navigation to the various display sections are not intuitive to get to - a page design sort of thing. Once you discover them they are useful, though.

Problems Reported by Curators:

- . . . the internal links [in a captured site] all led back to the original site, forcing us to use the Search feature to find the pages.
- My main concern is that I'm still having trouble viewing all of the files on the site . . . I'm confident that this will be resolved at some point. But I worry that it might break and break often. This doesn't make the tool stable.

3.3 Searching Files

Table 5 lists the file searching options in rank order by curators' ratings of "Very Useful". Generally, curators (*n*=14) conducted keyword-only searches and found these either very

useful (*n*=8) or somewhat useful (*n*=6). Many did not try searches limited either by file type (*n*=8; 47%) or by URL (*n*=10; 59%).

	N	**Very Useful**		**Somewhat Useful**		**Not Useful**		**Not Tried**	
		#	%	#	%	#	%	#	%
Keyword(s) only search	17	8	47%	6	35%	0	0%	3	18%
File type limited (type:) keyword search	17	6	35%	3	18%	0	0%	8	47%
URL limited (URL:) keyword search	17	4	24%	2	12%	1	6%	10	59%

Table 5. Usefulness of File Search Options

Comments from Curator:

- I honestly rarely used the search feature when reviewing my results. First, I'd check the Robot, Host, Response Codes and Mimetype reports to get an idea of the 'health' and extent of the site - I wanted to see if there were any exclusions, how many 404 Not Found errors, and then just how many pdfs, MSWord, html, image files the site included. My main concern at this point is to make sure that the crawl is capturing the stuff I want - reports, publications, etc. (which are reflected as pdf, html, txt, MSWord, Excel) and getting it into the system. Once I checked the reports, then I went to the files tab and displayed results by file type to review what was captured and add it to a collection.

3.4 Results: Mapping and Comparing

Curators were fairly evenly divided regarding their estimate of the usefulness of the map view of hosts from which site content was captured. A total of five curators each rated the map view as very useful, somewhat useful, and not useful (Table 6). Two curators were not sure of its usefulness.

N=17		**#**	**%**
	Very Useful	5	29%
	Somewhat Useful	5	29%
	Not Useful	5	29%
	Not Sure	2	12%

Table 6. Usefulness of Host Map View in Evaluating Results

Fourteen curators attempted to compare the results of two different captures of the same site. Most (*n*=11; 65%) had no difficultly doing this; however, three curators did experience difficulties.

Problems Reported by Curators:

- A usability issue I ran into - I kept forgetting to click the Capture link when viewing my captures, so sometimes I got lost when I wanted to compare results. Also, I tried the Compare link under captures a few times and it didn't work - I got an error screen. I did try again a different day it worked so stability of this feature is a concern.

- One curator wanted to compare the actual display of the home pages of two different captures of the same site but "could not keep open at the same time two different captures to compare the look and feel of the home page". This curator found this to be "a major fault".

Comments from Curators:

- The compare results feature is a major asset and tool - perhaps a backbone feature for WAS.

3.5 Recommendations and Considerations

Recommendations from Curators:

- I would like to be able to delete sites as I review a capture, i.e. the 404s. (Note: Perhaps the curator meant "files" instead of "sites"?)
- Add the ability to exclude sites or file types when comparing captures.

Consideration:

- Investigate and resolve any inconsistencies between the Overview Report and the captured content reported by a curator in Section 3.1.
- Investigate and resolve the issue of internal links in captured sites taking curators out of the archive to live web sites.
- Consider providing curators with a feature to easily display the captured content of their seed URLs (i.e., the "home page") as well as the ability to display different captures of the same seed URL.

4 Collections

4.1 Adding and Removing Content

Most curators (*n*=13; 76%) did not have any difficulty creating a collection. In all, 15 of 17 curators reported successfully adding entire site captures to their collections. Only three curators successfully added files to their collections, while four curators were unable to do so and 10 others did not attempt to do so (Table 7).

Overall, curators did not attempt to remove content, either entire sites or files, from their collections (Table 7). Given that curators were creating collections for the first time and were not directed to remove content from collections, this result seems reasonable.

Entire sites (N=17)		**Added**		**Removed**	
		#	**%**	**#**	**%**
	Yes	15	88%	5	29%
	No	0	0%	0	0%
	Not Tried	2	12%	12	71%
Individual files (N=17)		**Added**		**Removed**	
		#	**%**	**#**	**%**
	Yes	3	18%	2	12%
	No	4	24%	0	0%
	Not Tried	10	59%	15	88%

Table 7. Adding and Removing Collection Content

4.2 Searching Collections

Table 8 lists the searching options for collections in rank order by curators' ratings of "Very Useful". Similarly to their results when searching files (Section 3.3), most curators (*n*=11) conducted keyword-only searches and found these either very useful (*n*=6) or somewhat useful (*n*=5). Most did not try searches limited either by file type (*n*=8; 50%) or by URL (*n*=9; 56%).

	N	Very Useful		Somewhat Useful		Not Useful		Not Tried	
		#	**%**	**#**	**%**	**#**	**%**	**#**	**%**
Keyword(s) only search	17	6	35%	5	29%	3	18%	3	18%
File type limited (type:) search	16	4	25%	3	19%	1	6%	8	50%
URL limited (URL:) search	16	1	6%	4	25%	2	13%	9	56%

Table 8. Usefulness of Collection Search Options

Differently from their usefulness ratings of searching files, a greater number, although still small (3 versus 0), did not find searching collections useful. This may be related to some curators' reports that they were unable to search the content of their collections and received error messages related to WAS accessing the index used for searching and listing files. The WAS Release 4 Help Manual did state that the "Search and Files screens take about an hour to update after you have added or removed content from your collection". It is possible that curators did not wait a sufficient amount of time for their captured files to be indexed; however, it seems curators assumed the error message they received was indicative of a system problem. Their comments below illustrate their experience.

Problems Reported by Curators:

- Searching my collection for the term 'municipal code' returned an error screen that said something about Error finding the index. I tried limiting it to html, but achieved the same result.
- It doesn't look like the keyword search is working right now. I tried one collection with 4 captures in to using a generic word - water, or California. I only got results from [1] website.
- The ability to Search a collection or see the files in a collection was broken when we tried it. 'There was an error getting the index for searching or listing files. This index may be very large and is taking a long time with its initial loading.'
- collection searching did not work -- got 'Error getting index' msg for all searches
- I ran into a problem with my captured sites . . . The site was captured but wasn't being indexed . . .

4.3 Considerations

Considerations:

- The indexing delay between adding/removing collection content and being able to search that content currently returns an error message. If this delay cannot be eliminated, then perhaps changing this message to alert curators that their collection content is currently being indexed and will be available later is advisable. Additionally, is it possible to omit or "gray out" the Search and Files

tabs to serve as a visual indication when indexing is not complete? Alternately, is it feasible to eliminate the search field/function from the Search screen until the files are indexed and perhaps present text stating the sites files are being indexed and will be available for searching at a later time?

5 Overall Reactions

Overall, curators were very satisfied with this WAS release (Table 9), as one curator commented: "It shows continuous improvement over previous releases." Two curators were not satisfied with the release and one of these commented: "A lot of the new features (such as searching collections) would be really nice if they worked." Not surprisingly, when individual curators encountered major problems, their satisfaction with the WAS lowered. However, this release was generally very well received as reflected in the satisfaction rating of "very satisfied" by 65% (*n*=11) of the curators.

N=17		#	%
	Very Satisfied	11	65%
	Somewhat Satisfied	4	24%
	Not Satisfied	2	12%
	No Response	0	0%

Table 9. Overall Curator Satisfaction

5.1 WAS Help

Most curators thought the documentation provided was very helpful in answering questions about features and functions of the WAS (Table 10). In particular, all curators referenced the side bar information and 76% (*n*=13) found the information very helpful. No curators reported that the documentation provided was not helpful.

	N	Very Helpful		Somewhat Helpful		Not Helpful		Not Used	
		#	%	#	%	#	%	#	%
Side Bar Information	17	13	76%	4	24%	0	0%	0	0%
Detailed Guides / User Manual	17	10	59%	4	24%	0	0%	3	18%
Contextual Help	17	11	65%	3	18%	0	0%	3	18%

Table 10. Helpfulness of Documentation

5.2 What Curators Liked the Most

1. Improvements
   - Help screens
   - Workflow
   - Speed of captures
   - Displays
   - Performance
2. Bookmarklet
3. RSS feed
4. Compare results

5. Ease of use
   - Navigation
   - Adding content to collections
   - Capturing sites
   - Making changes to sites and collections

6. Ability to create collections

5.3 Areas for Improvement

1. Navigation & interface
   - "While viewing captured content, it's impossible to tell when the links lead you OUT of the archive and back to the original site."
   - "'Manage Collections' area still seems non-intuitive to me. The icons help make the actions more explicit, but the first thing I see when I click on 'manage collections' is still the 'create new collection' option, which instantly makes me think, 'but I want to manage, not create a new collection!' . . . I think intuitively, I'm looking to the left and below the 'Manage Collections' heading for actions."
   - "The navigation between areas really needs improvement. It relies on the 'discovery' method and should be to be thought through more."
   - "The process of capturing, editing captures, and viewing captures is still not easy for the uninitiated. . . . once you've got the hang of it there's not a problem."
   - "Breaking down the walls between the different modules"

2. Access to other curators' archived sites
   - For specific collaborations with other curators
   - To add content captured by any curator to a collection
   - To determine if another curator has captured a site

3. Speed
   - Navigating between managing and viewing captures
   - When comparing captures

4. Content capture
   - "The entire content capture engine seems very buggy." (Note: See Section 3.1, *Problem Reported by Curator*, for details.)

5. Content display
   - Viewing all files from a captured site

6. Timing of trial
   - "My dissatisfaction comes from the timing of the release and communication about it. You picked the absolute worst time of year for those of us on the quarter system . . . Related to that, a very brief introduction (4 bullet points and not a 20 page manual) in the announcement about what we are testing on this release and how that fits in with the entire project would have been very useful."
   - "The timing of the release was as bad as it could possibly have been. Though already late the release should have been postponed until November."

7. Desirable Feature Enhancements
    - "Thumbnails of the websites would be a welcome addition to both the 'Manage Sites' and 'View Captures' areas. This site is so text-heavy . . . "
    - Ability to schedule capture jobs
        - Within a date range
        - Annually
    - Subject heading thesaurus
    - Capture option: Directory + 1 link
    - Direct Help on a screen to the relevant section within the general help document
    - Ability to view site captures as a tree structure with filenames
    - Field for recording selector's notes about a site
    - Explanation of errors
    - Option to "deal with" robot exclusions (Note: Curator did not specify what this would ideally entail.)
    - Ability to create a "perma-link" or "stable URL", similar to a "tinyurl bibpurl", for collections, individual files, and captures, so catalogs, websites, and email messages can include links to the archived collections and files
    - Ability to view all sites on a single screen with a scroll bar, in Manage Sites and elsewhere
    - Ability to exclude specific mimetypes from captures
    - Include an 'enter' or 'submit' button on initial login screen for curator to select after entering username/password. (Note: The delay when the WAS initially loads caused some confusion regarding whether or not the curator had correctly logged in after hitting the 'enter' key on the keyboard.)
    - Provide "one-stop data management" for collection development (e.g., "a cross between an OPAC, a traditional database, and something like the Archivists Toolkit") by adding a module for the pre-capture phase that includes:
        - Ability to add sites, not yet captured, to collections
        - Make "subjects" a repeatable field
        - Add a collections field for curator to identify collection(s) to which a site might be assigned
        - Search and sort list of collections, whether captured or saved (Note: Perhaps the curator meant "sites", not "collections"?)